Exploiting Metadata for Off-line Handwritten Documents: Modeling and Applications

Ph.D Candidacy: Jin Chen

Thesis Advisor: Daniel Lopresti

Computer Science & Engineering, Lehigh University, Bethlehem, PA, United States

Introduction

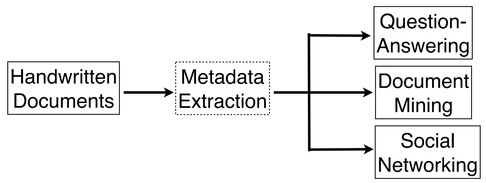
Document image analysis (DIA) is the subfield of digital image processing that aims at converting document images to a symbolic form for modification, storage, retrieval, reuse, and transmission [1]. However, more information is conveyed in addition to such a transcription, e.g., writer idiosyncrasies, data arrangement (tables, forms, etc.).

Metadata in Context [2]:

- *Descriptive Metadata:* describes a resource for discovery and identification, e.g., author, and writer idiosyncrasies.
- *Structural Metadata*: indicates how document components are organized, e.g., tables, pre-printed ruling lines, etc.

Thesis Hypothesis:

By exploiting various metadata in off-line handwritten documents, we are able to restore the original structure between documents, build new relationships from them, and facilitate problem solving in information retrieval tasks.



A high-level view of modules of document analysis system for various applications.

Table Detection in HW Documents

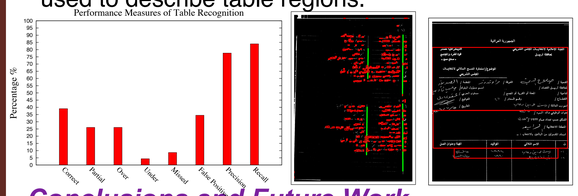
Table Detection via Dynamic Programming [3]

Each page is decomposed into:
1. one table
2. multiple tables

Row correlation is based on inside-space stream.

Table Detection Performance

- 62 Arabic handwritten documents are scanned at 600 dpi. The ground-truth for text regions (HW vs. MP) is available as bounding polygons.
- 20 pages for testing and the other 42 for training.
- Use area ratio-based measures proposed in Shafait and Smith [5], where bounding boxes are used to describe table regions.



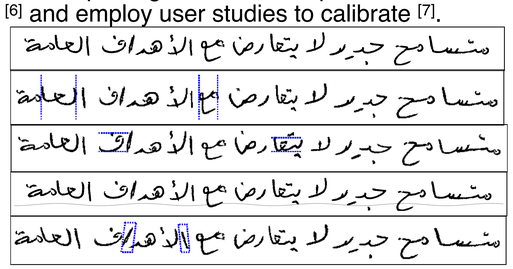
Conclusions and Future Work

- Existing methods in MP documents do not completely solve the problem in HW documents.
- Part of the errors are caused by complicated layouts, such as letter forms, signatures, etc.
- Future work includes detection methods requiring no row segmentation, and better layout analysis.

Writer ID w/ Severe Data Constraints

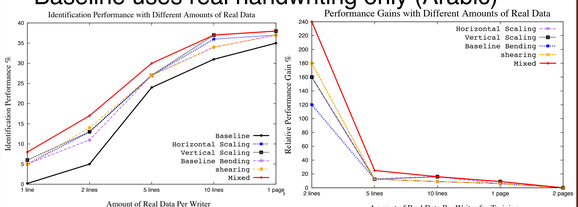
Scenarios and Methodologies

- Although usually assumed sufficient for research, data can become a problem for real-world writer ID, e.g., authors are unavailable, uncooperative.
- Model-Generated Handwriting and Model-Perturbed Handwriting can be used.
- We adopt Varga and Bunke's perturbation model [6] and employ user studies to calibrate [7].

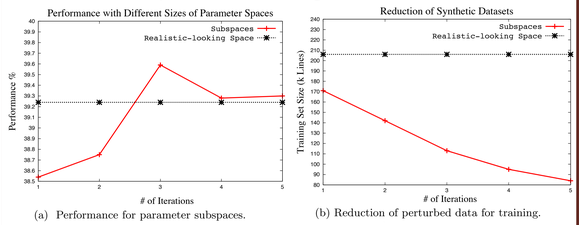


Writer Identification Performance

- Baseline uses real handwriting only (Arabic)



- Better Subspace Searching



(a) Performance for parameter subspaces. (b) Reduction of perturbed data for training.

Conclusions and Future Work

- MPH can be useful for writer ID, in addition to handwriting recognition.
- It is possible to find better subspace of parameters s.t. more gains are acquired.
- Future work includes simulation-based approach that requires no expensive classifier re-training.

Model-based Ruling Line Detection

Scenarios and Methodologies

- Pre-printed ruling lines are used to help people write neatly on paper. In document analysis, however, they create challenges for handwriting recognition and writer identification.
- Rewrite the *residual* function and solve it:

$$\begin{bmatrix} \Sigma 1 & \Sigma i & \Sigma x_{i,j} \\ \Sigma i & \Sigma i^2 & \Sigma i\ x_{i,j} \\ \Sigma x_{i,j} & \Sigma i\ x_{i,j} & \Sigma x_{i,j}^2 \end{bmatrix} \times \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \Sigma y_{i,j} \\ \Sigma i\ y_{i,j} \\ \Sigma x_{i,j}\ y_{i,j} \end{bmatrix}$$

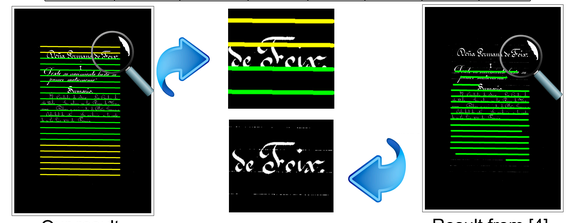
- Ground-truth on the page level and measure directly the detection errors as:

$$D[i] = M_{algorithm}[i] - M_{ground-truth}[i]$$

Ruling Line Detection Performance

Table 2. Performance comparison of our model-based ruling line detection and an existing algorithm [4].

Stats.	$\mathcal{P}.x$	$\mathcal{P}.y$	$\mathcal{L}$	$\mathcal{K}$	β_1	β_2	$\mathcal{H}$
Synthesis Dataset							
μ	-4.16	0.55	6.9	0	0	-0.01	0
σ	5.60	2.18	5.11	0	0.01	0.05	1.41
Germana Dataset							
μ	-12.30	3.80	12.85	0.25	-0.49	-0.09	0.25
σ	39.01	20.70	34.03	1.25	2.24	0.26	0.22
Germana Dataset (an existing method [4])							
μ	-49.40	57.60	70.40	-1.20	-2.40	0.02	0
σ	62.64	113.30	62.12	5.68	8.82	1.09	0



Conclusions and Future Work

- Propose an effective model-based ruling line detection algorithm.
- Scale-up experiments and the effects on applications such as writer ID are ongoing.

A Short Bio

Jin Chen has finished his fourth year at Lehigh University pursuing a Ph.D degree. He joined the Pattern Recognition Lab in Fall 2007 working with Professor Daniel Lopresti since then, on on-line/off-line handwriting related problems. He expects to graduate by May 2012.

References

1. G. Nagy, "Twenty years of document image analysis in pami," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 36–62, 2000.
2. NISO, "Understanding metadata," NISO Press, ISBN 1-880124-62-9.
3. J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, "Medium- independent table detection," in Document Recognition and Retrieval VIII (IS&T/SPIE Electronic Imaging, 2001, pp. 44– 55.
4. Y. Zheng, H. Li, and D. Doermann, "A parallel-line detection algorithm based on HMM decoding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 777–792, 2005.
5. F. Shafait and R. Smith, "Table detection in heterogeneous documents," in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 2010, pp. 65–72.
6. Varga, T. and Bunke, H., "Generation of synthetic training data for an hmm-based handwriting recognition system," in Proceedings of the seventh International Conference on Document Analysis and Recognition, 618–622 (2003).
7. Cheng, W. and Lopresti, D., "Parameter calibration for synthesizing realistic-looking variability in offline handwriting," in Proc. Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging), (2011).